Belayer: Modeling discrete and continuous spatial variation in gene expression from spatially resolved transcriptomics

Cong Ma[*1], Uthsav Chitra[*1], Shirley Zhang[1], and Benjamin J. Raphael[†1]

[1]Department of Computer Science, Princeton University

Abstract

Spatially resolved transcriptomics (SRT) technologies simultaneously measure gene expression and spatial location of cells in a 2D tissue slice, enabling the study of spatial patterns of gene expression. Current approaches to model spatial variation in gene expression assume either that gene expression is determined by discrete cell types or that gene expression varies continuously across a tissue slice. However, neither of these modeling assumptions adequately represent spatial variation in gene expression: the first assumption ignores continuous variation within a spatial region containing cells of the same type, while the second assumption does not allow for discontinuous changes in expression, e.g., due to a sharp change in cell type composition. We propose a model of spatial patterns in gene expression that incorporates both discontinuous and continuous spatial variation in gene expression. Specifically, we model the expression of a gene in a *layered tissue* slice as a piecewise linear function of a single spatial coordinate with potential discontinuities at tissue layer boundaries. We formulate the problem of inferring tissue layer boundaries and gene expression functions for all genes simultaneously. We derive a dynamic programming algorithm to find the optimal boundaries of tissue layers when these boundaries are lines parallel to a coordinate axis. We generalize this algorithm to arbitrarily curved tissue layers by transforming the tissue geometry to be axis-aligned using the theory of conformal maps from complex analysis. We implement these algorithms in a method called Belayer. Applying Belayer to simulated data and to spatial transcriptomics data from the human dorsolateral prefrontal cortex, we demonstrate that Belayer achieves both higher accuracy in clustering tissue layers compared to state-of-the-art SRT clustering methods, and higher accuracy in identifying cortical layer marker genes compared to commonly-used methods for identification of spatially varying genes.

Keywords: *spatially resolved transcriptomics, spatial variation, gene expression, layered tissues, segmented regression, conformal maps*

[*]These authors contributed equally and author order was decided by a coin toss.

[†]Correspondence: braphael@princeton.edu

1 Introduction

Spatially resolved transcriptomics (SRT) technologies simultaneously measure both gene expression and spatial location of cells in a two-dimensional tissue slice [48]. Examples of SRT technologies include in-situ hybridization (ISH) techniques such as MERFISH [17] and seqFISH+ [24], which are based on imaging fluorescence probes, and sequencing techniques such as Slide-seq [49], Slide-seqV2 [52] and the 10X Genomics Visium spatial transcriptomics platform [1, 51], which sequence barcoded mRNA molecules whose barcodes record both the spatial locations of the molecules and unique molecular identifiers (UMI) that indicate the count of mRNA molecules. SRT technologies have enabled a wide range of biological analyses, including analyses on the spatial organization of different cell types [40, 8, 38, 11, 29, 23, 6, 14, 42, 22, 53, 31, 13] and inter-cellular communication [65, 15, 20, 35].

Analysis of SRT data presents multiple challenges. First, most SRT technologies produce high-dimensional data; for example, sequencing-based technologies such as Slide-Seq [49] and spatial transcriptomics [51] measure the expression of 10,000 - 20,000 human genes across thousands or tens of thousands of spatial locations. Second, gene expression measurements from many current SRT technologies are sparse, with some genes not measured even if they are expressed. For example, some SRT technologies measure only hundreds of genes, while current sequencing based SRT technologies have very low UMI counts per spot; e.g., in the 10X Genomics Visium platform [1] each spot has a median of about 5,000 UMI counts, while Slide-seqV2 [52] reports median UMI counts of approximately 500. Finally, jointly modeling both gene expression and the spatial location of cells in a tissue requires an appropriate model of spatial variation in gene expression, and is thus more involved than applying existing models developed for bulk or single-cell gene expression data.

Current computational models for analyzing spatial variation in gene expression in SRT data make one of two distinct modeling assumptions. The first modeling assumption is that gene expression is determined by discrete cell types; specifically, there are a small number of clusters with distinct cell type compositions, and gene expression at a spot depends only on the cluster label, i.e. the cell type composition present at the spot. This modeling assumption is made by methods for identifying cell type clusters [69, 20, 47, 18]. These methods implicitly address data sparsity by sharing information across nearby spots, e.g., through the use of models such as hidden Markov random fields (HMRFs). Under this modeling assumption, large differences in gene expression between clusters are allowed, which corresponds to discrete shifts in cell type composition within a tissue slice. However, methods that make this modeling assumption also assume that gene expression is constant within each cluster, and thus do not account for continuous spatial variation of gene expression within a cluster.

The second modeling assumption is that gene expression varies continuously across a tissue slice. This assumption is usually made by methods that identify spatially varying genes [55, 54, 21, 70] or that construct low-dimensional representations [60, 57]. For example, SpatialDE [55] and SPARK [54] model gene expression with a Gaussian Process in which the covariance of expression between a pair of spots decreases as their spatial distance increases. This modeling assumption is justified by the fact that cellular environments continuously vary across space, and thus gene expression is affected by cellular environments and inter-cellular communications [34, 7, 26]. However, most methods that make this modeling assumption do not account for large discrete changes in cell type composition across the tissue and the consequent discontinuous change in expression.

We introduce a method called Belayer to analyze spatially resolved transcriptomics data using a *global* model of tissue organization and an explicit definition of gene expression that combines both discrete and continuous variation in space. Specifically, we define *layered tissues*, a global model of tissue organization, that consist of consecutive layers of cell types. Layered tissues are common in many organs, e.g., human skin consists of three layers [32], the cerebral cortex consists of six layers [39], and the retina has 10 distinct layers of neurons [44]. In the simplest case, a layered tissue has a one-dimensional spatial structure, and we

model the expression of a gene using a *piecewise continuous* function of a single spatial variable. *Piecewise* functions allow for discontinuities in expression where there are sharp changes in cell type composition in space, such as between tissue layers, while *continuous* functions model gradients of gene expression within a tissue layer, e.g. [43, 27]. To reduce overfitting with sparse SRT data, we model gene expression using *piecewise linear* functions, which are specified by a small number of parameters. The inference of piecewise linear gene expression functions in 1D is related to changepoint detection [5] and segmented regression [2, 10, 63], well-studied problems in time-series analysis whose maximum likelihood solutions can be computed using dynamic programming. We extend the classical dynamic programming algorithm for segmented regression to jointly infer piecewise linear gene expressions for all genes simultaneously. We also demonstrate that dimensionality reduction using generalized PCA (GLM-PCA) [58] preserves the one-dimensional structure of a layered tissue, thus formalizing the ad hoc dimensionality reduction steps often made when analyzing SRT data. Next, we generalize this approach to two-dimensional layered tissues by using tools from complex analysis, namely conformal maps — complex analytic functions that locally preserve angles between curves [45] — to transform a general 2D layered tissue into a layered tissue with a one-dimensional structure. We then extend our dynamic programming algorithm to more general 2D layered tissues.

We implement our algorithms in a method called Belayer and we apply Belayer to simulated data and to 10X Visium spatial transcriptomics data from the human dorsolateral prefrontal cortex [40]. We demonstrate that Belayer achieves both higher accuracy in clustering tissue layers compared to state-of-the-art SRT clustering methods, and higher accuracy in identifying cortical layer marker genes compared to commonly-used methods for identification of spatially varying genes.

2 Layered tissues and expression functions

Modeling spatial patterns of gene expression is complicated by the large variability in the spatial structure of tissues — e.g. some parts of brain tissues have a layered structure while muscle tissues have a striated structure [4] — as well as the sparsity of the gene expression profiles produced by current spatially resolved transcriptomics (SRT) technologies. To avoid overfitting the data, it is helpful to make simplifying assumptions about the spatial structure of the tissue slice T and/or the spatial patterns of gene expression. Here, motivated by the layered structure of the skin, brain, eyes, and other organs [4], we focus on *layered tissue slices* (Figure 1A) which we define as follows.

Definition 1. *An* L-layered tissue slice *is a region* $T \subseteq \mathbb{R}^2$ *containing* $L-1$ *non-intersecting smooth curves* $\Gamma_1, \ldots, \Gamma_{L-1}$*, or* layer boundaries*, satisfying:*

1. *each curve* Γ_ℓ *has end points on the boundary* ∂T *of* T*;*
2. *every point* $p \in T$ *is contained in a region* R *bounded by* ∂T *and* at most two *curves* Γ_ℓ *and* $\Gamma_{\ell'}$*.*

The $L-1$ layer boundaries $\Gamma_1, \ldots, \Gamma_{L-1}$ partition the tissue slice T into L regions $R_1, \ldots, R_L$, or *layers*. The layers $R_1, \ldots, R_L$ of a L-layered tissue slice T represent biologically distinct regions in the tissue slice T. For example, in a tissue slice T from the skin, R_1, R_2, R_3 may represent the epidermis, dermis, and hypodermis layers, where each layer consists of unique cell types and has unique functions [32]. Without loss of generality we assume that the layers $R_1, \ldots, R_L$ and layer boundaries $\Gamma_1, \ldots, \Gamma_{L-1}$ are ordered so that layer R_1 is bounded by tissue boundary ∂T and layer boundary Γ_1, layer R_ℓ is bounded by layer boundaries $\Gamma_{\ell-1}$ and Γ_ℓ for $\ell = 2, \ldots, L-1$, and layer R_L is bounded by layer boundary Γ_{L-1} and tissue boundary ∂T.

We describe the spatial distribution of the expression of a gene g in a tissue T with an *expression function* $f_g : T \to \mathbb{R}$, where $f_g(x, y)$ measures the expression of gene g at spatial location $(x, y) \in T$.

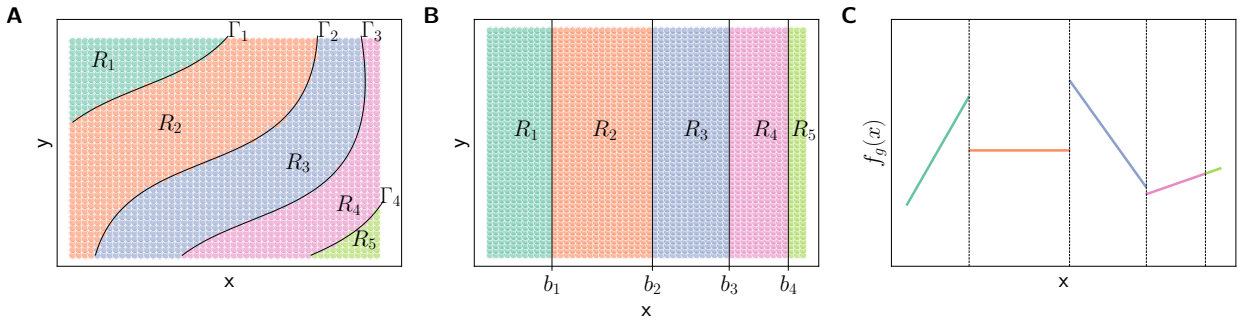

Figure 1: Layered tissue slices and piecewise linear expression functions. **(A)** A 5-layered tissue slice T with layers $R_1, \ldots, R_5$ and layer boundaries $\Gamma_1, \ldots, \Gamma_4$. **(B)** An axis-aligned 5-layered tissue slice with layers $R_1, \ldots, R_5$ and layer boundaries $x = b_1, \ldots, x = b_4$. **(C)** Expression function $f_g(x)$ for gene g along the x coordinate for the axis-aligned layered tissue slice in (B). For all panels, colors indicate layers and black lines indicate layer boundaries.

For example, a gene g whose expression is uniform across the tissue has a constant expression function $f_g(x, y) = c$, while a marker gene g for a specific region $R \subseteq T$ has the expression function $f_g(x, y) = c \cdot 1_{\{(x,y) \in R\}} + c' \cdot 1_{\{(x,y) \notin R\}}$.

More generally, if T is a L-layered tissue slice with layers $R_1, \ldots, R_L$ and the expression of gene g in layer R_ℓ is given by layer-specific expression function $f_{g,\ell}(x, y)$, then gene g has the expression function $f_g(x, y) = \sum_{\ell=1}^{L} f_{g,\ell}(x, y) \cdot 1_{\{(x,y) \in R_\ell\}}$. We assume that the layer-specific expression functions $f_{g,\ell}(x, y)$ are continuous functions, and thus $f_g(x, y)$ is a piecewise-continuous function with discontinuities allowed at the layer boundaries Γ_ℓ. These discontinuities correspond to differences in expression due to changes in cell type composition between different layers R_ℓ of the tissue T. In the next section, we describe how one models SRT data using the expression functions $f_g(x, y)$.

3 Axis-aligned layered tissues

We begin by studying the simplest L-layered tissue slice: the *axis-aligned* L-layered tissue slice, where each layer boundary Γ_ℓ is a line $x = b_\ell$ parallel to the y-axis (Figure 1B). We assume that the expression of a gene g at position (x, y) depends only on the distance from the layer boundaries $x = b_\ell$. Under this assumption, the layer-specific expression functions $f_{g,\ell}(x, y)$ are functions of only the x-coordinate: $f_{g,\ell}(x, y) = f_{g,\ell}(x)$. Thus, the expression function $f_g(x, y)$ also only depends on the x-coordinate, i.e. $f_g(x) = \sum_{\ell=1}^{L} f_{g,\ell}(x) \cdot 1_{\{b_{\ell-1} < x \leq b_\ell\}}$ where for convenience we define $b_0 = -\infty$ and $b_L = \infty$. We call each b_ℓ a *breakpoint* of the expression function f_g.

The simplest model for $f_g(x)$ is a piecewise constant function, which corresponds to each gene g having a constant expression in each layer R_ℓ. Such a piecewise constant model is implicit in methods that assume constant expression in contiguous regions of cell types; e.g. methods that use hidden Markov random field (HMRF) models [69, 64, 20, 56]. We generalize these approaches by modeling continuous variation in expression within a layer; e.g. due to gradients of gene expression [43, 27, 12, 16]. Since current SRT technologies have limited dynamic range and spatial resolution, inference of complicated expression functions may be prone to overfitting. Thus, we use the simplest form of continuous, non-constant spatial variation within each layer and model each layer-wise expression function $f_{g,\ell}(x)$ as a *linear function*. Then each expression function $f_g(x)$ is a piecewise linear function (Figure 1C) with shared breakpoints

$b_1, \ldots, b_{L-1}$ across all genes corresponding to the axis-aligned layer boundaries $\Gamma_1, \ldots, \Gamma_{L-1}$, i.e.

$$f_g(x) = \sum_{\ell=1}^{L} (\beta_{g,\ell} x + \alpha_{g,\ell}) \cdot 1_{\{b_{\ell-1} < x \leq b_\ell\}}, \tag{1}$$

where $\beta_{g,\ell}$ and $\alpha_{g,\ell}$ are the slope and y-intercept, respectively, of gene g in layer R_ℓ. We define $\mathcal{L}(b_1, \ldots, b_{L-1})$ to be the set of piecewise linear functions $f(x)$ with breakpoints $b_1, \ldots, b_{L-1}$, and we define $\mathcal{L}$ to be the set of linear functions $f(x)$.

We contrast our model for the expression functions f_g with another commonly used model for expression functions, Gaussian Processes (GPs) [55, 54, 57, 36]. In this approach each expression function $f_g(x, y) \sim \mathcal{GP}\big(0, k(\mathbf{s}, \mathbf{s}')\big)$ is an independent sample from a GP with mean function 0 and covariance function $k(\mathbf{s}, \mathbf{s}')$ between spatial locations $\mathbf{s} = (x, y)$ and $\mathbf{s}' = (x', y')$. The covariance functions $k(\mathbf{s}, \mathbf{s}')$ used by existing methods, including SpatialDE [55] and SPARK [54], are continuous and thus do not model piecewise continuous expression functions $f_g(x)$. We show in the Appendix that it is possible to model piecewise linear functions with GPs using a one-dimensional blockwise covariance function; however, to our knowledge such covariance functions have not been used to model spatially resolved transcriptomics data.

3.1 Axis-Aligned L-Layered Problem

We aim to infer the layer boundaries $x = b_\ell$ and expression functions $f_g(x)$ that maximize the likelihood of the observed SRT data. We represent SRT data as an expression matrix $\mathbf{A} = [a_{i,g}] \in \mathbb{R}^{N \times G}$ and a spatial location matrix $\mathbf{S} \in \mathbb{R}^{2 \times N}$ with column $\mathbf{s}_i = (x_i, y_i)$ indicating the spatial location of spot i. We define the Axis-Aligned L-Layered Problem as the the following maximum likelihood estimation problem.

Axis-Aligned L-Layered Problem. *Given SRT data $(\mathbf{A}, \mathbf{S})$ and a number L of layers, find layer boundaries $x = b_1, \ldots, x = b_{L-1}$ and piecewise linear expression functions $f_1, \ldots, f_G \in \mathcal{L}(b_1, \ldots, b_{L-1})$ that maximize the log-likelihood of the data:*

$$\underset{\substack{b_1 < b_2 < \cdots < b_{L-1} \\ f_1, \ldots, f_G \in \mathcal{L}(b_1, \ldots, b_{L-1})}}{\arg\max} \sum_{g=1}^{G} \left(\sum_{i=1}^{N} \log P\big(a_{i,g} \mid f_g(x_i)\big) \right). \tag{2}$$

When there is $L = 1$ layer, each expression function $f_g(x)$ is a linear function, and thus the maximum log-likelihood (2) in the Axis-Aligned L-Layered Problem reduces to

$$\max_{f_1, \ldots, f_G \in \mathcal{L}} \sum_{g=1}^{G} \left(\sum_{i=1}^{N} \log P\big(a_{i,g} \mid f_g(x_i)\big) \right) = \sum_{g=1}^{G} \max_{f_g \in \mathcal{L}} \left(\sum_{i=1}^{N} \log P\big(a_{i,g} \mid f_g(x_i)\big) \right). \tag{3}$$

The maximization on the right-hand side of (3) is a linear regression problem of finding the linear function f_g that best fits the observed data. Thus, (3) is solved by computing a separate linear regression for each gene $g = 1, \ldots, G$.

More generally, when each expression function $f_g(x) = \sum_{\ell=1}^{L} f_{g,\ell}(x) \cdot 1_{\{b_{\ell-1} < x \leq b_\ell\}}$ is an L-piecewise linear function with *known* breakpoints $b_1, \ldots, b_{L-1}$, then the maximum log-likelihood (2) in the Axis-Aligned L-Layered Problem reduces to

$$\max_{f_1, \ldots, f_G \in \mathcal{L}(b_1, \ldots, b_{L-1})} \sum_{g=1}^{G} \left(\sum_{i=1}^{N} \log P\big(a_{i,g} \mid f_g(x_i)\big) \right) = \sum_{g=1}^{G} \sum_{\ell=1}^{L} \left(\max_{f_{g,\ell} \in \mathcal{L}} \sum_{i : b_{\ell-1} < x_i \leq b_\ell} \log P\big(a_{i,g} \mid f_{g,\ell}(x_i)\big) \right), \tag{4}$$

Then, each expression function f_g is computed by solving L separate regression problems, one regression problem for each pair $x = b_{\ell-1}, x = b_\ell$ of consecutive layer boundaries. On the other hand, when the breakpoints $b_1, \ldots, b_{L-1}$ of the piecewise linear expression functions f_g are *unknown*, one will have to compute a regression over *all* pairs of possible layer boundaries for each gene g, as we describe in the next section.

We emphasize that the regression approach described above can be used with different probability distributions $P\big(a_{i,g} \mid f_g(x_i)\big)$ for the observed counts $a_{i,g}$. In particular, following [58, 50, 46] we model UMI counts $a_{i,g}$ with a Poisson distribution $a_{i,g} \overset{\text{i.i.d.}}{\sim} \text{Poisson}\Big(C_i \cdot \exp\big(f_g(x_i)\big)\Big)$ where C_i is the total UMI count at spot i. Using the Poisson model we solve the regression problems in (4) with Poisson regression. Another alternative is to model normalized expression values $a_{i,g}$, with a Gaussian distribution $a_{i,g} \overset{\text{i.i.d.}}{\sim} N(f_g(x_i), \sigma^2)$ where σ^2 is a shared variance parameter. Under the Gaussian model, the regression problems in (4) are linear regressions.

Visualization and binning of SRT data. To simplify the visualization of SRT data $(\mathbf{A}, \mathbf{S})$ from an axis-aligned tissue slice, we combine expression values for a gene from spots with the same x-coordinate into a single "binned" expression value (Figure 2A). Specifically, for each gene g we construct a binned expression value $\widetilde{a}_{j,g}$ that estimates the gene expression for all spots with the same x-coordinate $\widetilde{x}_j$. For the Poisson model the binned expression $\widetilde{a}_{j,g} = \sum_{i:x_i=\widetilde{x}_j} a_{i,g}$ is the sum of the UMI counts $a_{i,g}$ at spots $\mathbf{s}_i = (x_i, y_i)$ with x-coordinate $x_i = \widetilde{x}_j$, while for the Gaussian model the binned expression $\widetilde{a}_{j,g} = \frac{1}{|\{i:x_i=\widetilde{x}_j\}|} \sum_{i:x_i=\widetilde{x}_j} a_{i,g}$ is the average expression. We show in the Appendix that binning the data does not affect inference of the expression functions $f_1, \ldots, f_G$ in the Axis-Aligned L-Layered Problem, as the expression functions obtained by maximizing the log-likelihood (2) with the binned expressions $[\widetilde{a}_{j,g}]$ are equal to the expression functions obtained by maximizing (2) with the original expressions $[a_{i,g}]$.

In the applications presented below, it is often the case that there are spots $\mathbf{s}_i = (x_i, y_i)$ with *approximately* equal x-coordinates x_i. Thus, we extend the binning approach to construct binned expressions $\widetilde{a}_{j,g}$ from spots $\mathbf{s}_i = (x_i, y_i)$ with x-coordinates $x_i \approx \widetilde{x}_j$. Further, for UMI count data we plot the *normalized* binned expressions $\log(\widetilde{a}_{i,g}/\widetilde{C}_i)$, where $\widetilde{C}_i$ is the sum of the total UMI counts C_i for all spots i in bin j (Figure 2A). These normalized counts have the same scale as the expression functions f_g. See the Appendix for more details.

3.2 A dynamic programming algorithm for the Axis-Aligned L-Layered Problem

When there is only one gene, i.e. $G = 1$, then the Axis-Aligned L-Layered Problem is also known as *segmented regression*. Segmented regression is a classical problem in statistics and time-series analysis and has a well-known dynamic programming (DP) solution [9, 2, 10, 63]. Here, we present a DP algorithm for the Axis-Aligned L-Layered Problem that extends the classical DP algorithm for segmented regression [2, 10, 63] to any number G of genes.

Briefly, the DP algorithm is as follows. Let $M_{n,\ell}$ be the maximum log-likelihood of the first n datapoints fit to ℓ-piecewise linear expression functions $f_1, \ldots, f_G$, i.e.

$$M_{n,\ell} = \max_{\substack{b_1 < b_2 < \cdots < b_{\ell-1} \\ f_1, \ldots, f_G \in \mathcal{L}(b_1, \ldots, b_{\ell-1})}} \sum_{g=1}^{G} \left(\sum_{i=1}^{n} \log P\big(a_{i,g} \mid f_g(x_i)\big) \right), \tag{5}$$

where we assume without loss of generality that the spots are ordered by x-coordinate so that $x_1 \leq x_2 \leq \cdots \leq x_N$.

The best piecewise linear fit for the first n datapoints with ℓ pieces corresponds to: (1) the best piecewise linear fit for the first n' datapoints with $\ell - 1$ pieces and (2) the best linear function fit for the remaining

$n - n'$ datapoints, for some $n' < n$. This yields the following recurrence:

$$M_{n,\ell} = \max_{n'<n} \left[M_{n',\ell-1} + \sum_{g=1}^{G} \left(\max_{f_g \in \mathcal{L}} \sum_{i=n'}^{n} \log P\big(a_{i,g} \mid f_g(x_i)\big) \right) \right], \tag{6}$$

where the inner maximization is solved using regression.

The DP algorithm consists of using the recursion (6) to fill in a DP table column-by-column followed by a pass backwards through the table to identify the L-piecewise linear expression functions $f_1, \ldots, f_G$ and breakpoints $b_1, \ldots, b_{L-1}$. The run-time of the DP algorithm is upper-bounded as $O(LN^2G \cdot P_0)$, where P_0 is the runtime to solve the inner maximization in (6) for a single gene g. For the case where the gene expression values $a_{i,g}$ follow the Gaussian model, the run-time can be shortened to $O(LN^2G)$ by using linear algebra techniques from [2]. When N is large, it is also possible to reduce the run-time of the dynamic programming algorithm – at the expense of spatial resolution – by restricting the coordinates n', n in the recurrence (6) to a subsequence of the N datapoints.

4 Modeling L-layered tissues with conformal maps

For an arbitrary L-layered tissue slice with layer boundaries $\Gamma_1, \ldots \Gamma_{L-1}$, we also assume that the expression of a gene g depends only on the distance to the layer boundaries Γ_ℓ. However, because the layer boundaries Γ_ℓ are not necessarily axis-aligned, one cannot immediately write the expression function f_g as a function of a single variable. An elegant solution is obtained by using a conformal map, a tool from complex analysis [3]. A conformal map $\Phi : \mathbb{C} \to \mathbb{C}$ is a complex function that locally preserves angles; equivalently, Φ is conformal if it is analytic and has non-zero derivative everywhere. Note that the inverse Φ^{-1} of a conformal map Φ is also conformal. Conformal maps are used to solve differential equations in the plane $\mathbb{R}^2$ with complicated boundary conditions – e.g. heat flow in a plate or airflow around a wing – by identifying $\mathbb{R}^2$ with the complex plane $\mathbb{C}$ [45].

Here, we use conformal maps to transform a layered tissue slice T into an axis-aligned layered tissue slice T'. Ideally, we would derive a conformal map $\Phi : T \to T'$ that transforms tissue slice T to axis-aligned tissue slice T', with the constraints that Φ maps each layer boundary Γ_ℓ in T to the corresponding layer boundary $x = b_\ell$ in axis-aligned tissue slice T'. Unfortunately, constructing a conformal map satisfying these additional contraints is challenging, as the standard constructions of conformal maps only allow constraints on the boundary of T [45]. In general, it is unclear whether a conformal map satisfying additional constraints on non-boundary curves always exists.

As an alternative, we derive L conformal maps $\Phi_\ell : R_\ell \to R'_\ell$ for $\ell = 1, \ldots, L$, where Φ_ℓ transforms layer R_ℓ of tissue slice T to the corresponding layer R'_ℓ of axis-aligned tissue slice T'. We require that each conformal map Φ_ℓ maps the layer boundaries $\Gamma_{\ell-1}, \Gamma_\ell$ of layer R_ℓ to the corresponding layer boundaries $x = b_{\ell-1}, x = b_\ell$ of layer R'_ℓ, respectively, which is equivalent to the following constraints on the conformal map Φ_ℓ:

$$\mathbf{Re}\, \Phi_\ell(\Gamma_{\ell-1}) = b_{\ell-1} \text{ and } \mathbf{Re}\, \Phi_\ell(\Gamma_\ell) = b_\ell, \tag{7}$$

where $\mathbf{Re}$ denotes the real part of a complex number. We define $\mathcal{C}_T(b_{\ell-1}, b_\ell)$ as the set of conformal maps Φ_ℓ that satisfy (7). Notice that (7) are constraints on the boundary of each layer R_ℓ, and thus such maps are guaranteed to exist. Then, the map $\Phi : T \to T'$ from tissue slice T to axis-aligned tissue slice T' is given by the piecewise sum of the L conformal maps $\Phi_1, \ldots, \Phi_L$, that is

$$\Phi(x, y) = \sum_{\ell=1}^{L} \Phi_\ell(x, y) \cdot \mathbf{1}_{(x,y) \in R_\ell}. \tag{8}$$

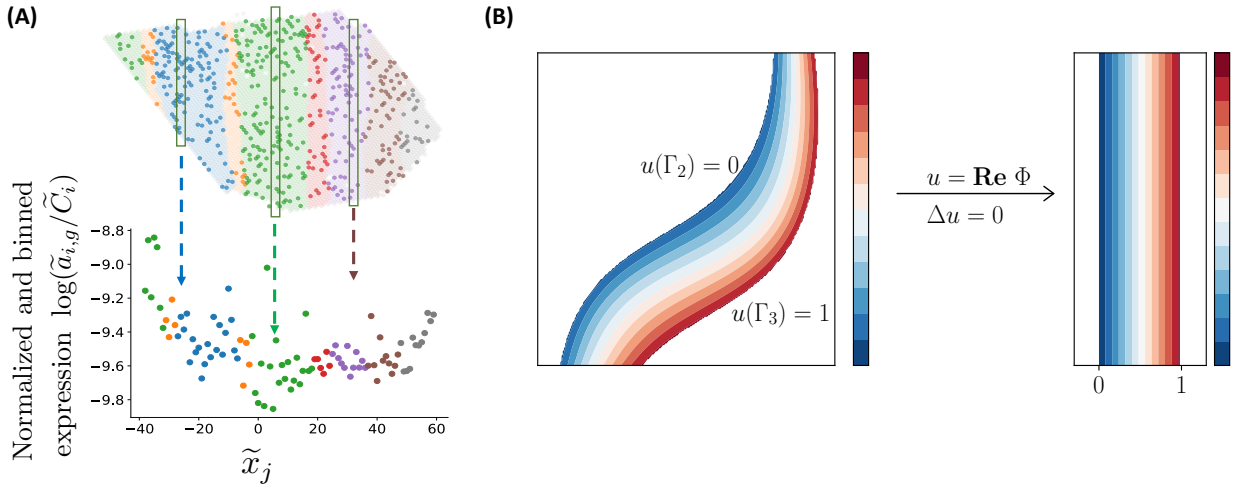

Figure 2: **(A)** (Top) Spatially resolved transcriptomics data from an axis-aligned layered tissue slice with non-zero UMIs indicated in darker color. (Bottom) Visualization of 1D expression values obtained from binning and normalization of UMI counts from 2D spots with similar x-coordinates. **(B)** A conformal map Φ maps a curved layer to a vertical strip aligned along x coordinate. The real part $u = \mathbf{Re}\,\Phi$ is a harmonic function and satisfies heat equation $\Delta u = 0$. The color shows contours of the heat u in the layered tissue.

By the symmetry principle [45], if the limits of Φ_ℓ and $\Phi_{\ell+1}$ at Γ_ℓ are the same and are continuous for all ℓ, then Φ is a conformal map. However, it is unknown whether these equality and continuity conditions are guaranteed to hold for arbitrary layer boundaries $\Gamma_1, \ldots, \Gamma_L$.

We model the expression of gene g in axis-aligned tissue slice T' using a piecewise linear function $f_g(x)$ as in Section 3. Then, our goal is to find conformal maps $\Phi_1, \ldots, \Phi_L$ and piecewise linear functions $f_g(x)$ such that the measured gene expression $\mathbf{A}$ at spots $\mathbf{S}$ in the tissue slice T are best fit by piecewise linear functions $f_g(x)$ in the corresponding axis-aligned tissue T'. We define the L-Layered Problem as the following maximum likelihood estimation problem.

L-Layered Problem. *Given SRT data $(\mathbf{A}, \mathbf{S})$, layered tissue slice $T \subseteq \mathbb{C}$, and a number L of pieces, find an axis-aligned tissue slice T' with layers $x = b_1, \ldots, x = b_{L-1}$, conformal maps $\Phi_\ell \in \mathcal{C}_T(b_{\ell-1}, b_\ell)$, and L-piecewise linear expression functions $f_1, \ldots, f_G \in \mathcal{L}(b_1, \ldots, b_{L-1})$ that maximize the log-likelihood of the data:*

$$\underset{\substack{b_1<b_2<\cdots<b_{L-1} \\ \Phi_\ell \in \mathcal{C}_T(b_{\ell-1},b_\ell) \text{ for } \ell=1,\ldots,L \\ f_1,\ldots,f_G \in \mathcal{L}(b_1,\ldots,b_{L-1})}}{\arg\max} \sum_{g=1}^{G} \left(\sum_{i=1}^{N} \log P\Big(a_{i,g} \mid f_g\big(\Phi(x_i, y_i)\big)\Big) \right). \tag{9}$$

The L-Layered Problem generalizes the assumption in the Axis-Aligned L-Layered Problem that the expression function $f_g(x, y)$ of a gene g at position (x, y) depends only on the distance from the layer boundaries $x = b_{\ell-1}$ and $x = b_\ell$; in this more general formulation, the expression function $f_g(x, y)$ of gene g is constant along contours that "interpolate" between adjacent layer boundaries $\Gamma_{\ell-1}$ and Γ_ℓ (Figure 2B).

In general, solving the L-Layered Problem without additional constraints on the conformal maps $\Phi_1, \ldots, \Phi_L$ is challenging. Below, we give algorithms for solving two special cases of L-Layered Problem which provide useful approximations on real data: (1) when the layer boundaries $\Gamma_1, \ldots, \Gamma_{L-1}$ are known and (2) when the layer boundaries $\Gamma_1, \ldots, \Gamma_{L-1}$ are lines.

4.1 Approximate layer boundaries Γ_ℓ are given

Suppose we are given prior information about the spatial organization of a tissue slice T in the form of approximate layer boundaries $\widetilde{\Gamma}_\ell$. While these approximate layer boundaries may not directly correspond to the true layer boundaries Γ_ℓ, they can nevertheless be used to compute conformal maps Φ_ℓ. From these conformal maps, we can then estimate piecewise linear expression functions $f_1, \ldots, f_G$, but without requiring that the breakpoints b_ℓ of the piecewise linear functions f_g occur at the mapped approximate layer boundaries $\Phi_\ell(\Gamma_\ell)$. In this way, we can incorporate prior information about tissue geometry (e.g. from prior knowledge of the tissue structure or from H&E images) without requiring that the expression functions conform exactly to this prior information.

Specifically, suppose we are given approximate layer boundaries $\widetilde{\Gamma}_1, \ldots, \widetilde{\Gamma}_{L'-1}$, where L' is not necessarily equal to the desired number L of layer boundaries. We constrain each conformal map Φ_ℓ so that the layer boundaries $\widetilde{\Gamma}_{\ell-1}$ and $\widetilde{\Gamma}_\ell$ are mapped to lines $x = \widetilde{b}_{\ell-1}$ and $x = \widetilde{b}_\ell$, respectively, for some choice of real numbers $\widetilde{b}_1, \ldots, \widetilde{b}_{L'-1}$. These constraints on the conformal maps Φ_ℓ are equivalent to the constraints $\mathbf{Re}\,\Phi_\ell(\widetilde{\Gamma}_{\ell-1}) = \widetilde{b}_{\ell-1}$ and $\mathbf{Re}\,\Phi_\ell(\widetilde{\Gamma}_\ell) = \widetilde{b}_\ell$ for all $\ell = 1, \ldots, L'$, which is in turn equivalent to the constraint $\Phi_\ell \in \mathcal{C}_{\widetilde{T}}(\widetilde{b}_{\ell-1}, \widetilde{b}_\ell)$, where $\widetilde{T}$ is tissue slice T equipped with the approximate layer boundaries $\widetilde{\Gamma}_1, \ldots, \widetilde{\Gamma}_{L'-1}$ and approximate latyers $\widetilde{R}_1, \ldots, \widetilde{R}_{L'-1}$. We define the conformal map $\Phi = \sum_{\ell=1}^{L'} \Phi_\ell(x, y) \cdot \mathbf{1}_{(x,y)\in\widetilde{R}_\ell}$ as the piecewise sum of the L' conformal masp Φ_ℓ. In practice, we choose $\widetilde{b}_1, \ldots, \widetilde{b}_{L'-1}$ so that $\widetilde{b}_\ell - \widetilde{b}_{\ell-1} = d(\widetilde{\Gamma}_\ell, \widetilde{\Gamma}_{\ell-1})$ where $d(\cdot, \cdot)$ is the partial Hausdorff distance [28]; see the Appendix for details.

Thus, given approximate layer boundaries $\widetilde{\Gamma}_1, \ldots, \widetilde{\Gamma}_{L'-1}$ and lines $x = \widetilde{b}_1, \ldots, x = \widetilde{b}_{L'-1}$, we solve the following optimization problem which we call the *L-Layered Problem with Approximate Layer Boundaries*:

$$\underset{\substack{b_1<b_2<\cdots<b_{L-1} \\ \Phi_\ell\in\mathcal{C}_{\widetilde{T}}(\widetilde{b}_{\ell-1},\widetilde{b}_\ell) \text{ for } \ell=1,\ldots,L' \\ f_1,\ldots,f_G\in\mathcal{L}(b_1,\ldots,b_{L-1})}}{\arg\max} \sum_{g=1}^{G} \left(\sum_{i=1}^{N} \log P\Big(a_{i,g} \mid f_g\big(\Phi(x_i, y_i)\big)\Big) \right). \tag{10}$$

The conformal maps Φ_ℓ in (10) can be solved for separately from the breakpoints b_ℓ and expression functions f_g by using the observation that the constraint $\Phi_\ell \in \mathcal{C}_{\widetilde{T}}(\widetilde{b}_{\ell-1}, \widetilde{b}_\ell)$ *uniquely* defines the real part of the conformal map Φ_ℓ. This is because, in the constraints $\mathbf{Re}\,\Phi_\ell(\widetilde{\Gamma}_{\ell-1}) = \widetilde{b}_{\ell-1}$ and $\mathbf{Re}\,\Phi_\ell(\widetilde{\Gamma}_\ell) = \widetilde{b}_\ell$, the real part $\mathbf{Re}\,\Phi_\ell$ of each conformal map Φ_ℓ is a harmonic function, and thus can be written as the solution to the heat equation with specified boundary conditions. That is, if one fixes curve $\widetilde{\Gamma}_{\ell-1}$ to have constant temperature $\widetilde{b}_{\ell-1}$ and $\widetilde{\Gamma}_\ell$ to have constant temperature $\widetilde{b}_\ell$, then $\mathbf{Re}\,\Phi_\ell(x, y)$ is the temperature at point (x, y) (Figure 2B). In practice, because there is no closed-form solution to the continuous heat equation for arbitrarily shaped boundaries, we solve the heat equation on a discretized tissue slice using a random walk-based approach [33, 25]. See the Appendix for more details.

After computing the conformal maps $\Phi_1, \ldots, \Phi_{L'}$, the optimization problem in (10) reduces to

$$\underset{\substack{b_1<b_2<\cdots<b_{L-1} \\ f_1,\ldots,f_G\in\mathcal{L}(b_1,\ldots,b_{L-1})}}{\arg\max} \sum_{g=1}^{G} \left(\sum_{i=1}^{N} \log P\Big(a_{i,g} \mid f_g\big(\Phi(x_i, y_i)\big)\Big) \right). \tag{11}$$

This is an instance of the Axis-Aligned L-Layered Problem with transcript counts $a_{i,g}$ and transformed spots $\Phi(\mathbf{s}_i)$. Thus, we solve (11) by using the dynamic programming algorithm for the Axis-Aligned L-Layered Problem (Section 3.2).

4.2 Layer boundaries Γ_ℓ are lines

The second special case of the L-Layered Problem is when the layer boundaries $\Gamma_1, \ldots, \Gamma_{L-1}$ are lines. Let $\mathcal{Q}(T)$ be the set of lines Γ with endpoints on the boundary ∂T of the tissue T. In this case, given

lines $x = b_1, \ldots, x = b_{L-1}$, the L-Layered Problem reduces to solving the following optimization problem which we call the *Linear L-Layered Problem*:

$$\underset{\Gamma_1,\ldots,\Gamma_{L-1}\in\mathcal{Q}(T)}{\arg\max} \sum_{\ell=1}^{L} c(R_\ell), \tag{12}$$

$$\text{where } c(R_\ell) = \max_{\substack{f_1,\ldots,f_G\in\mathcal{L}(b_1,\ldots,b_{L-1})\\ \Phi_\ell\in\mathcal{C}_T(b_{\ell-1},b_\ell)}} \sum_{i:(x_i,y_i)\in R_\ell} \sum_{g=1}^{G} \log P\Big(a_{i,g} \mid f_g\big(\Phi_\ell(x_i,y_i)\big)\Big). \tag{13}$$

We derive a dynamic programming algorithm to solve (12) for any convex tissue slice T and any function $c(R)$ that maps subsets $R \subseteq T$ of the tissue to $\mathbb{R}$. We note that the specific function $c(R)$ in (13) can be computed by solving the heat equation as described in Section 4.1. Our dynamic programming algorithm generalizes the dynamic programming algorithm for the Axis-Aligned L-Layered Problem from Section 3.2.

Briefly, the dynamic programming algorithm for solving (12) is as follows. Without loss of generality, let $\mathbf{s}_1, \ldots, \mathbf{s}_{N_{\text{boundary}}}$ be the spots on the boundary ∂T of the tissue T in clockwise order. For convenience, we define $[n, m]$ be the sequence $(n+1, \ldots, m-1)$ of indices when $m > n$, and the sequence $(n+1, \ldots, N_{\text{boundary}}, 1, \ldots, m-1)$ when $n > m$. We define $T_{n,m} \subseteq T$ to be the region of the tissue T that is formed by drawing a line Γ between spots $\mathbf{s}_n$ and $\mathbf{s}_m$ and has boundary spots $\{\mathbf{s}_i\}_{i\in(n,m)}$.

Let $M_{n,m,\ell}$ be the best fit with ℓ layers (with $\ell - 1$ linear layer boundaries) in the region $T_{n,m}$, i.e.

$$M_{n,m,\ell} = \max_{\Gamma_1,\ldots,\Gamma_{\ell-1}\in\mathcal{Q}(T_{n,m})} \sum_{\ell'=1}^{\ell} c(R_{\ell'}). \tag{14}$$

Then, the best fit with ℓ layers in the region $T_{n,m}$ can be decomposed to the sum of: (1) the best fit with $\ell - 1$ layers in the region $T_{n',m'}$ and (2) the fit $c(T_{n,m} \setminus T_{n',m'})$ for the region $T_{n,m} \setminus T_{n',m'}$, for some $n', m' \in [n, m]$. Thus, we have the following recurrence relationship:

$$M_{n,m,\ell} = \max_{n',m'\in[n,m]} \left(M_{n',m',\ell-1} + c(T_{n,m} \setminus T_{n',m'})\right). \tag{15}$$

After computing the values $M_{n,m,\ell}$ for all $1 \le n, m \le N_{\text{boundary}}$ and $\ell = 1, \ldots, L-1$, the optimal value of (12) is then derived as

$$\max_{\Gamma_1,\ldots,\Gamma_{L-1}\in\mathcal{Q}(T)} \sum_{\ell=1}^{L} c(R_\ell) = \max_{n,m} \left(M_{n,m,L-1} + M_{m,n,1}\right). \tag{16}$$

The dynamic programming algorithm consists of using the recursion (15) to fill in a table, followed by computing (16) to identify the best fit with $L - 1$ linear layer boundaries for the entire tissue T. See the Appendix for more details.

4.2.1 Layer boundaries Γ_ℓ are parallel lines

We also derive a more computationally efficient algorithm for the case when the layer boundaries $\Gamma_1, \ldots, \Gamma_{L-1}$ are *parallel* lines. In this case, each conformal map Φ_ℓ is a rotation map $P_\theta : \mathbb{R}^2 \to \mathbb{R}^2$ with the same angle θ, where $P_\theta(x, y)$ rotates points (x, y) about the origin by angle $\theta \in [0, \pi)$. Then, the L-Layered Problem reduces to the following optimization problem which we call the *θ-Rotated L-Layered Problem*:

$$\underset{\substack{\theta\in[0,\pi)\\ b_1<b_2<\cdots<b_{L-1}\\ (f_1,\ldots,f_G)\in\mathcal{L}(b_1,\ldots,b_{L-1})}}{\arg\max} \sum_{g=1}^{G}\sum_{i=1}^{N} \log P\Big(a_{i,g} \mid f_g\big(P_\theta(x_i,y_i)\big)\Big). \tag{17}$$

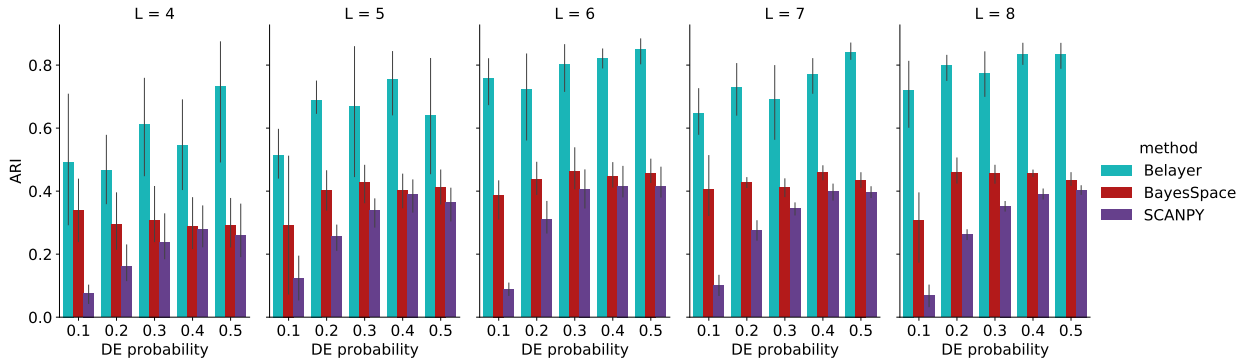

Figure 3: Comparison of Belayer, BayesSpace, and SCANPY in identifying spatially distinct cell clusters in simulated spatially resolved transcriptomics data from layered tissues. Performance of each method is evaluated according to the Adjusted Rand Index (ARI) and shown for different values of the number L of layers and differential expression (DE) probability. Error bars indicate variation from 5 randomly simulated datasets for each parameter setting.

For a fixed angle θ, the optimization problem in (17) reduces to the Axis-Aligned L-Layered Problem with expression matrix A and rotated spots $P_\theta(x_i, y_i)$. Thus, we solve the θ-Rotated L-Layered Problem by performing a parameter sweep over angles $\theta \in [0, \pi)$ and finding the value of θ that maximizes the objective of (17).

5 Improving run-time with dimensionality reduction

The algorithms described above for solving different versions of the L-Layered Problem solve a separate regression problem for each gene. This can result in large run-times when the number of genes G is large, particularly when using a Poisson model for the UMI counts $a_{i,g}$. This is because Poisson regression requires solving a convex optimization problem, which is generally slower than linear regression.

To reduce the runtime for SRT data $(\mathbf{A}, \mathbf{S})$, we use generalized principal components analysis (GLM-PCA) [58] to produce a low-dimensional representation of the expression matrix $\mathbf{A}$, and then run the algorithms presented above on the top-$2L$ generalized principal components $u_1, \ldots, u_{2L} \in \mathbb{R}^N$ using the Gaussian model. This results in large reductions in run-time: for example, for the Axis-Aligned L-Layered Problem, we reduce the run-time of our DP algorithm from $O(LN^2G \cdot P_0)$ under the Poisson model to $O(LN^2 \cdot 2L) = O(N^2L^2)$, corresponding to the run-time for DP with the Gaussian model and $G = 2L$ genes. After computing the breakpoints $b_1, \ldots, b_L$, we then estimate the expression functions f_g for each gene g by computing L separate Poisson regressions as described in Section 3.1, for which the run-time is only $O(LGP_0)$.

Our use of GLM-PCA is motivated by the observation that for SRT data $(\mathbf{A}, \mathbf{S})$ following the Poisson model with piecewise linear expression functions f_g and breakpoints $b_1, \ldots, b_{L-1}$, the top-$2L$ generalized principal components $u_1, \ldots, u_{2L}$ of the expression matrix $\mathbf{A}$ are approximately piecewise linear with breakpoints $b_1, \ldots, b_{L-1}$. See the Appendix for more details.

6 Results

We implemented the algorithms described above for solving the L-Layered Problem in a software package called Belayer, available at github.com/raphael-group/belayer. We evaluated Belayer on simulated SRT data

and on SRT data from the human dosorlateral prefrontal cortex (DLPFC) [40].

6.1 Simulated data

We evaluated Belayer on simulated SRT data $(\mathbf{A}, \mathbf{S})$ from a rectangular tissue slice T with parallel layer boundaries lines $\Gamma_1, \ldots, \Gamma_{L-1}$ that are rotated by an angle θ from the x-axis, where θ is uniformly randomly chosen from $[-\frac{\pi}{4}, \frac{\pi}{4}]$. The observed spots $\mathbf{S}$ are equally spaced in the tissue slice T and form a 70×50 grid. We generated UMI count $a_{i,g}$ using Splatter [66] with $G = 1000$ genes such that: the rotated expression function $f_g(P_\theta(x, y))$ is piecewise linear in each layer for each gene g (using the the trajectory simulation feature of Splatter), the median UMI count per spot is around 1100, and the probability that a gene has a different linear expression function across layers, which we call DE probability, is set to a fixed value (using the differential expression probability parameter in Splatter). We generated 5 simulated SRT datasets $(\mathbf{A}, \mathbf{S})$ for each value of $L = 4 \ldots, 8$ and each value of DE probability in $\{0.1, 0, 2, 0.3, 0.4, 0.5\}$.

We ran Belayer to solve the θ-Rotated L-Layered Problem (Section 4.2.1), and compared Belayer to two other approaches: BayesSpace [69] and the SCANPY [62] implementation of leiden clustering algorithm [59]. Note that BayesSpace uses the spatial information while SCANPY does not. (We do not compare against stLearn [47] as it requires an H&E image as input. See the Appendix for details on data pre-processing and for parameter choices for each method.) We assume all methods know the true number L of layers. We assessed the performance of each method by computing the Adjusted Rand Index (ARI) between the L layers/clusters estimated by each method and the true layers. We found that Belayer outperformed BayesSpace and SCANPY clustering in identifying layers across all choices of simulation parameters (Figure 3), demonstrating that by modeling the layered tissue structure and continuous gradients in expression, Belayer is able to more accurately identify clusters with distinct spatial expression patterns in a tissue slice.

6.2 Real data

We next evaluated Belayer on spatial transcriptomics data from the human dorsolateral prefrontal cortex (DLPFC) obtained using the 10X Visium technology [40]. This dataset consists of 12 DLPFC tissue slices from three donors. Each slice was manually annotated [40] into six layers and white matter (WM). A list of 128 marker genes for the DLPFC was obtained from [40] and previous analyses [41, 67].

6.2.1 Cortical layer identification

We first compared Belayer against BayesSpace [69], stLearn [47] and SCANPY [62] on the four DLPFC tissue slices from Donor 1. For these four tissue slices, the manually annotated layer boundaries are approximately parallel lines with $L = 9$ layers. Thus, we used Belayer to solve the θ-Rotated L-layered Problem (Section 4.2.1) and learn the L layers. We evaluated each method by computing the ARI between the manually annotated layer labels and the layers identified by the method. Note that in some samples, the manually annotated labels correspond to two layers separated in space. We counted these as separate layers and accordingly separated any clusters of BayesSpace and stLearn that are separated in space when computing ARIs. We found that Belayer has noticeably higher ARI compared to BayesSpace, stLearn, and SCANPY (Figure 4A), demonstrating that Belayer learns more accurately clusters by leveraging the layered structure of the tissue.

Next we compared the methods on the 8 DLPFC slices from Donors 2 and 3. The manually annotated layer boundaries of these slices are not axis-aligned. Thus, we used Belayer to solve the L-Layered Problem with Approximate Layer Boundaries (Section 4.1) using the manually annotated layer boundaries to construct the conformal map. We found that Belayer outperformed the other methods in the identification

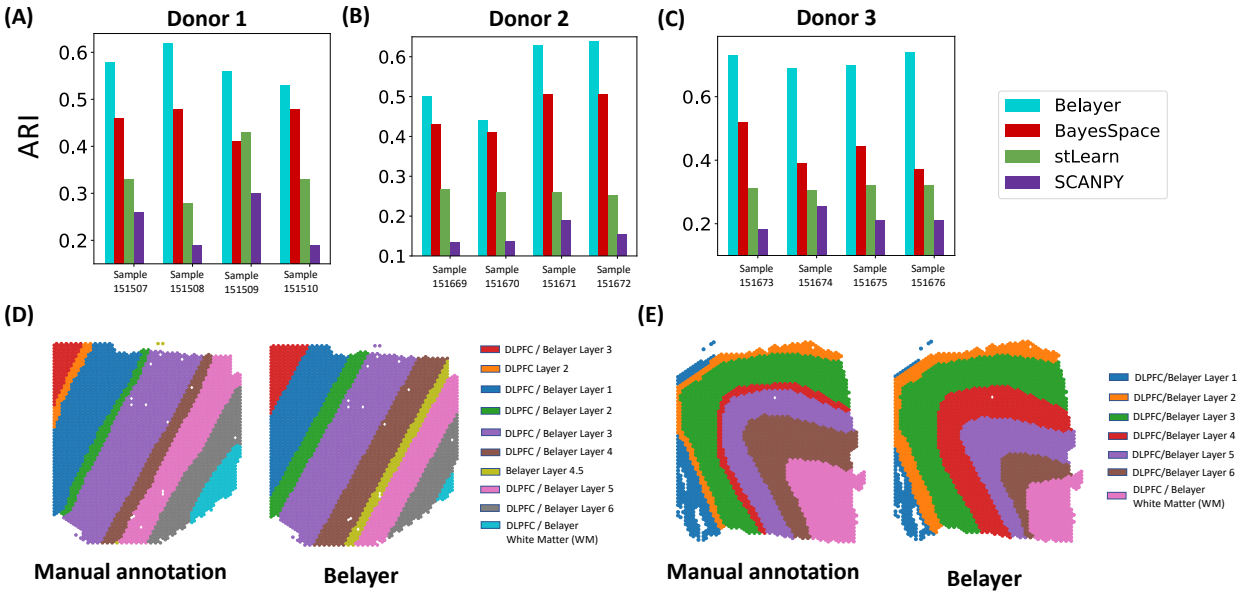

Figure 4: Comparison of Belayer, BayesSpace, stLearn, and SCANPY in identifying annotated layers in spatially resolved transcriptomics data from the dorsolateral prefrontal cortex (DLPFC) of three donors. **(A)** ARI for each method compared to manually annotated layers from each sample from Donor 1. Here, Belayer solves the θ-Rotated L-Layered Problem (Section 4.2.1). **(B), (C)** Same as (A) for Donors 2 and 3. Since the tissue slices from these donors are not axis-aligned, Belayer finds conformal maps (Section 4.1) to transform these tissue slices to axis-aligned tissue slices. **(D)** The layers identified by Belayer and the manually annotated layers for (axis-aligned) sample 151508 from Donor 1 and **(E)** sample 151673 from Donor 3. Layers identified by Belayer are labeled and colored according to the maximally overlapping layer from the manual annotation.

of layers (Figure 4B-C), suggesting that our piecewise linear assumption on the expression functions is appropriate. However, we emphasize that this comparison is overly generous to Belayer since Belayer uses information from the manually annotated layer boundaries to construct the conformal maps.

6.2.2 Identifying spatially varying genes

Finally, we compared the genes with layer-specific expression patterns inferred by Belayer to the spatially varying genes identified by other methods. We derived a list of such layer-specific genes for DLPFC tissue slice sample "151508" (shown in Figure 4D) from the expression functions $\widehat{f}_g$ inferred by Belayer as follows. We ran Belayer using $L = 6$ as it achieves the best ARI compared to the manually annotated layers. After excluding genes with low UMI counts (genes where more than 75% of binned spots had fewer than 10 UMIs), we ranked genes according to their largest layer-specific slope $\max_{\ell=1\ldots L} |\widehat{\beta}_{g,\ell}|$ across the $L = 6$ layers $\widehat{R}_1, \ldots, \widehat{R}_L$ identified by Belayer. We compared the overlap in rankings between these genes and known cortical marker genes from [40, 41, 67], and performed the same comparison for ranked lists of spatially varying genes from SpatialDE [55] and SPARK [54], two well-known methods for identifying spatially varying genes in spatially resolved transcriptomics data. We found that Belayer achieved higher AUPRC (0.041) compared to ranking genes according to the p-values of spatial variation computed by SpatialDE (0.02) and SPARK (0.037) (Figure 5A). We emphasize that because the list of known marker genes is small ($< 1\%$ of all $G \approx 15000$ genes), all three methods have low AUPRC.

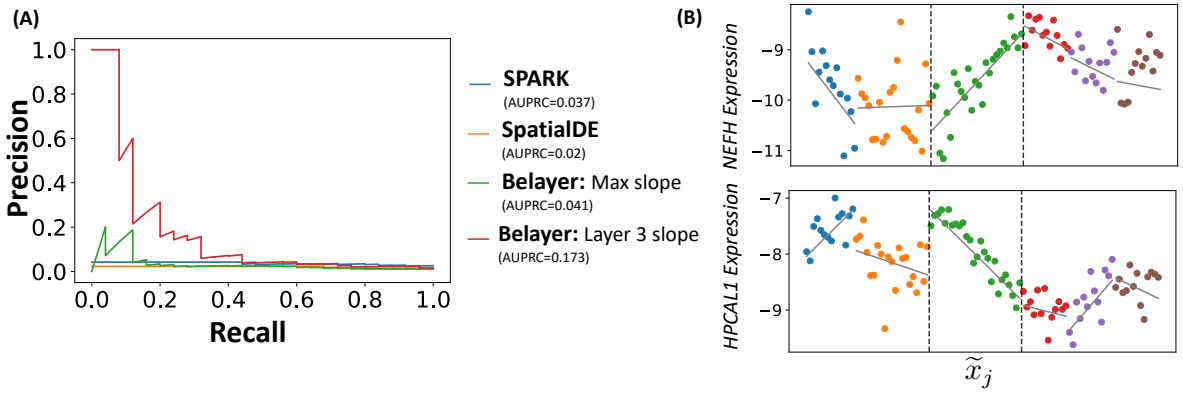

Figure 5: **(A)** Precision-recall curves for identifying marker genes in an axis-aligned DLPFC tissue slice (sample 151508) using four different methods. "Belayer: Max slope" corresponds to ranking genes by $\max_{\ell=1,\ldots,L} |\widehat{\beta}_{g,l}|$ and "Belayer: Layer 3 slope" corresponds to ranking genes by $|\widehat{\beta}_{g,3}|$ **(B)** Expression functions $f_g(x)$ learned by Belayer for genes *NEFH* and *HPCAL1* that have large slopes $|\beta_{g,3}|$ in the third layer.

We also observed that some layers are more predictive of marker genes than others. For example, ranking genes by the magnitude of their slope $|\widehat{\beta}_{g,3}|$ in the third layer identified by Belayer resulted in a much larger AUPRC (0.17) compared to scoring genes by their maximum slope $\max_{\ell=1,\ldots,L} |\widehat{\beta}_{g,\ell}|$ across all layers (Figure 5A). The genes g with large slope $|\widehat{\beta}_{g,3}|$ in the third layer are also biologically interesting, and we highlight two specific genes in Figure 5B. *NEFM* — the gene with the largest slope $|\widehat{\beta}_{g,3}|$ in the third layer — is a known cortical marker gene and is also a biomarker for neuronal damage [30]. On the other hand, *HPCAL1* — the gene with the fourth largest slope $|\widehat{\beta}_{g,3}|$ in the third layer — is not an annotated cortical marker gene but is reported to be involved in neuronal signalling [68, 61]. Our results demonstrate that incorporating layer-specific variation is important for identifying spatially varying genes, and suggest that the slopes $|\widehat{\beta}_{g,\ell}|$ identified by Belayer are a useful criteria for identifying cortical marker genes.

Finally, we note that the third cortical layer identified by Belayer is a combination of two manually annotated layers. Ranking genes g by their slope $|\widehat{\beta}_{g,\ell}|$ in either one of the manually annotated layers ℓ, has smaller AUPRC compared to ranking genes by their slope in the third layer identified by Belayer. This suggests that while the layers identified by Belayer do not exactly match the manually annotated layers, they potentially correspond to other biologically relevant partitions of the tissue.

7 Discussion

We introduce a new method, Belayer, to analyze spatial variation in gene expression from spatially resolved transcriptomics (SRT) data from *layered tissues*. Belayer models the expression of each gene with a *piecewise linear expression function* whose value depends only on the distance from the layer boundaries of the tissue. This piecewise linear model allows for both discrete changes in expression between layers – e.g. due to changes in cell type compositions – as well as continuous variation in gene expression within layers – e.g. due to gradients of gene expression. In the simplest case of an axis-aligned tissue structure, we infer the maximum likelihood expression function using a dynamic programming algorithm that is related to the classical problems of changepoint detection [5] and segmented regression [2, 10, 63]. We extend our approach to arbitrary layered tissues using the theory of conformal maps [45] from complex analysis [3], which are related to harmonic functions (e.g. the heat equation) and are often used to solve partial differential equations in engineering applications. We provide algorithms to solve two important special cases of the general

problem for specified layer boundaries and linear layer boundaries.

We show that Belayer outperforms existing approaches in identifying the spatial distribution of cell types and spatially varying genes in both simulated SRT data and SRT data from the human dorsolateral prefrontal cortex (DLPFC) [40]. Additionally, Belayer discovers genes with layer-specific and continuous spatial variation in expression, corresponding to both known cortical marker genes and genes with potentially novel cortical functions (Figure 5B). These results demonstrate that our piecewise linear model is a reasonable approach for the identification of tissue layers and the discovery of layer-specific marker genes.

There are a number of directions for future investigation. First, it would be desirable to provide an algorithm to solve (or approximately solve) the general L-Layered Problem. One possible approach is to extend the dynamic programming algorithm for lines to a larger class of layer boundaries. Second, it would helpful to further minimize overfitting by incorporating regularization and rigorous statistical testing into our segmented linear regression; e.g. using the Chow test [19] for change-point detection. Such extensions might also allow for *nonlinear* piecewise continuous expression functions, assuming the data has sufficient spatial resolution. Third, is to extend the definition of a layered tissue to account for more complex geometries, e.g. muscle tissues with concentric or striated layer structures [4]. These more complicated geometries can also be modeled with conformal maps, although the problem of inferring the conformal map from data becomes more complicated with tissue geometries that are less constrained than layered tissues. A fourth direction is to further investigate the expression functions learned by Belayer. In particular, the observation that the slopes of the expression functions learned by Belayer recovers known cortical layer marker genes better than existing methods inspires further study of the novel genes that are identified by our model, including genes with discontinuities and changes in slope at layer boundaries. Finally, it would be interesting to extend our approach to model other spatially resolved data, such as spatial proteomics [37].

8 Acknowledgements

U.C. is supported by NSF GRFP DGE 2039656. This research is supported by National Cancer Institute (NCI) grant U24CA264027 to B.J.R.

References

[1] 10x Genomics. Spatial transcriptomics, 2021.

[2] J. Acharya, I. Diakonikolas, J. Li, and L. Schmidt. Fast algorithms for segmented regression. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2878–2886, New York, New York, USA, 20–22 Jun 2016. PMLR.

[3] L. V. Ahlfors. *Complex Analysis*. McGraw-Hill Book Company, 2 edition.

[4] B. Alberts, D. Bray, K. Hopkin, A. D. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Essential cell biology*. Garland Science, 2015.

[5] S. Aminikhanghahi and D. J. Cook. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367, 2017.

[6] A. Andersson, J. Bergenstråhle, M. Asp, L. Bergenstråhle, A. Jurek, J. Fernández Navarro, and J. Lundeberg. Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Communications Biology*, 3(1):565, 2020.

[7] E. Armingol, A. Officer, O. Harismendy, and N. E. Lewis. Deciphering cell–cell interactions and communication from gene expression. *Nature Reviews Genetics*, 22(2):71–88, 2021.

[8] M. Asp, S. Giacomello, L. Larsson, C. Wu, D. Fürth, X. Qian, E. Wärdell, J. Custodio, J. Reimegård, F. Salmén, et al. A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. *Cell*, 179(7):1647–1660, 2019.

[9] R. Bacher, N. Leng, L.-F. Chu, Z. Ni, J. A. Thomson, C. Kendziorski, and R. Stewart. Trendy: segmented regression analysis of expression dynamics in high-throughput ordered profiling experiments. *BMC bioinformatics*, 19(1):380–380, 10 2018.

[10] J. Bai and P. Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78, 1998.

[11] E. Berglund, J. Maaskola, N. Schultz, S. Friedrich, M. Marklund, J. Bergenstråhle, F. Tarish, A. Tanoglidi, S. Vickovic, L. Larsson, et al. Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nature Communications*, 9(1):1–13, 2018.

[12] J. Briscoe and S. Small. Morphogen rules: design principles of gradient-mediated embryo patterning. *Development*, 142(23):3996–4009, 12 2015.

[13] D. M. Cable, E. Murray, V. Shanmugam, S. Zhang, M. Diao, H. Chen, E. Z. Macosko, R. A. Irizarry, and F. Chen. Cell type-specific differential expression in spatial transcriptomics. *bioRxiv*, 2021.

[14] D. M. Cable, E. Murray, L. S. Zou, A. Goeva, E. Z. Macosko, F. Chen, and R. A. Irizarry. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, 2021.

[15] Z. Cang and Q. Nie. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature Communications*, 11(1):1–13, 2020.

[16] M. S. Cembrowski, J. L. Bachman, L. Wang, K. Sugino, B. C. Shields, and N. Spruston. Spatial gene-expression gradients underlie prominent heterogeneity of ca1 pyramidal neurons. *Neuron*, 89(2):351–368, 2016.

[17] K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, and X. Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233), 2015.

[18] B. Chidester, T. Zhou, and J. Ma. Spicemix: Integrative single-cell spatial modeling for inferring cell identity. *bioRxiv*, 2021.

[19] G. C. Chow. Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28(3):591–605, 1960.

[20] R. Dries, Q. Zhu, R. Dong, C.-H. L. Eng, H. Li, K. Liu, Y. Fu, T. Zhao, A. Sarkar, F. Bao, R. E. George, N. Pierson, L. Cai, and G.-C. Yuan. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biology*, 22(1):78, 2021.

[21] D. Edsgärd, P. Johnsson, and R. Sandberg. Identification of spatial expression trends in single-cell gene expression data. *Nature Methods*, 15(5):339–342, 2018.

[22] M. Elosua-Bayes, P. Nieto, E. Mereu, I. Gut, and H. Heyn. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Research*, 49(9):e50–e50, 02 2021.

[23] R. Elyanow, R. Zeira, M. Land, and B. J. Raphael. STARCH: copy number and clone inference from spatial transcriptomics data. *Physical Biology*, 18(3):035001, 03 2021.

[24] C.-H. L. Eng, M. Lawson, Q. Zhu, R. Dries, N. Koulena, Y. Takei, J. Yun, C. Cronin, C. Karp, G.-C. Yuan, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, 568(7751):235–239, 2019.

[25] L. J. Grady and E. L. Schwartz. *Anisotropic interpolation on graphs: The combinatorial Dirichlet problem*. Citeseer, 2003.

[26] F. J. Grisanti Canozo, Z. Zuo, J. F. Martin, and M. A. H. Samee. Cell-type modeling in spatial transcriptomics data elucidates spatially variable colocalization and communication between cell-types in mouse brain. *Cell Systems*, 13(1):58–70.e5, 2022/01/30 2022.

[27] F. Hildebrandt, A. Andersson, S. Saarenpää, L. Larsson, N. Van Hul, S. Kanatani, J. Masek, E. Ellis, A. Barragan, A. Mollbrink, E. R. Andersson, J. Lundeberg, and J. Ankarklev. Spatial transcriptomics to define transcriptional patterns of zonation and structural components in the mouse liver. *Nature Communications*, 12(1):7046, 2021.

[28] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 1993.

[29] A. L. Ji, A. J. Rubin, K. Thrane, S. Jiang, D. L. Reynolds, R. M. Meyers, M. G. Guo, B. M. George, A. Mollbrink, J. Bergenstråhle, et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell*, 182(2):497–514, 2020.

[30] M. Khalil, C. E. Teunissen, M. Otto, F. Piehl, M. P. Sormani, T. Gattringer, C. Barro, L. Kappos, M. Comabella, F. Fazekas, A. Petzold, K. Blennow, H. Zetterberg, and J. Kuhle. Neurofilaments as biomarkers in neurological disorders. *Nature Reviews Neurology*, 14(10):577–589, 2018.

[31] V. Kleshchevnikov, A. Shmatko, E. Dann, A. Aivazidis, H. W. King, T. Li, R. Elmentaite, A. Lomakin, V. Kedlian, A. Gayoso, M. S. Jain, J. S. Park, L. Ramona, E. Tuck, A. Arutyunyan, R. Vento-Tormo, M. Gerstung, L. James, O. Stegle, and O. A. Bayraktar. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature Biotechnology*, 2022.

[32] P. A. Kolarsick, M. A. Kolarsick, and C. Goodwin. Anatomy and physiology of the skin. *Journal of the Dermatology Nurses' Association*, 3(4):203–213, 2011.

[33] G. F. Lawler. *Random walk and the heat equation*, volume 55. American Mathematical Soc., 2010.

[34] E. S. Lein, M. J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, A. F. Boe, M. S. Boguski, K. S. Brockway, E. J. Byrnes, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445(7124):168–176, 2007.

[35] D. Li, J. Ding, and Z. Bar-Joseph. Identifying signaling genes in spatial single-cell expression data. *Bioinformatics*, 37(7):968–975, 2021.

[36] Q. Li, M. Zhang, Y. Xie, and G. Xiao. Bayesian modeling of spatial molecular profiling data via Gaussian process. *Bioinformatics*, 37(22):4129–4136, 06 2021.

[37] E. Lundberg and G. H. H. Borner. Spatial proteomics: a powerful discovery tool for cell biology. *Nature Reviews Molecular Cell Biology*, 20(5):285–302, 2019.

[38] M. Mantri, G. J. Scuderi, R. Abedini-Nassab, M. F. Wang, D. McKellar, H. Shi, B. Grodner, J. T. Butcher, and I. De Vlaminck. Spatiotemporal single-cell RNA sequencing of developing chicken hearts identifies interplay between cellular differentiation and morphogenesis. *Nature Communications*, 12(1):1–13, 2021.

[39] F. Martini, M. J. Timmons, R. B. Tallitsch, W. C. Ober, C. W. Garrison, K. B. Welch, and R. T. Hutchings. *Human anatomy*. Pearson/Benjamin Cummings San Francisco, CA, 2006.

[40] K. R. Maynard, L. Collado-Torres, L. M. Weber, C. Uytingco, B. K. Barry, S. R. Williams, J. L. Catallini, M. N. Tran, Z. Besich, M. Tippani, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience*, 24(3):425–436, 2021.

[41] B. J. Molyneaux, P. Arlotta, J. R. Menezes, and J. D. Macklis. Neuronal subtype specification in the cerebral cortex. *Nature Reviews Neuroscience*, 8(6):427–437, 2007.

[42] R. Moncada, D. Barkley, F. Wagner, M. Chiodin, J. C. Devlin, M. Baron, C. H. Hajdu, D. M. Simeone, and I. Yanai. Integrating microarray-based spatial transcriptomics and single-cell rna-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature Biotechnology*, 38(3):333–342, 2020.

[43] A. E. Moor, Y. Harnik, S. Ben-Moshe, E. E. Massasa, M. Rozenberg, R. Eilam, K. B. Halpern, and S. Itzkovitz. Spatial reconstruction of single enterocytes uncovers broad zonation along the intestinal villus axis. *Cell*, 175(4):1156–1167, 2018.

[44] M. N and A. K. Y. Neuroanatomy, retina. StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing, 2021.

[45] Z. Nehari. *Conformal mapping*. Courier Corporation, 2012.

[46] L. Pachter. Models for transcript quantification from rna-seq, 2011.

[47] D. Pham, X. Tan, J. Xu, L. F. Grice, P. Y. Lam, A. Raghubar, J. Vukovic, M. J. Ruitenberg, and Q. Nguyen. stlearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *bioRxiv*, 2020.

[48] A. Rao, D. Barkley, G. S. França, and I. Yanai. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220, 2021.

[49] S. G. Rodriques, R. R. Stickels, A. Goeva, C. A. Martin, E. Murray, C. R. Vanderburg, J. Welch, L. M. Chen, F. Chen, and E. Z. Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019.

[50] A. Sarkar and M. Stephens. Separating measurement and expression models clarifies confusion in single-cell rna sequencing analysis. *Nature Genetics*, 53(6):770–777, 2021.

[51] P. L. Ståhl, F. Salmén, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

[52] R. R. Stickels, E. Murray, P. Kumar, J. Li, J. L. Marshall, D. J. Di Bella, P. Arlotta, E. Z. Macosko, and F. Chen. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nature Biotechnology*, 39(3):313–319, 2021.

[53] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, r. Mauck, William M, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, 06 2019.

[54] S. Sun, J. Zhu, and X. Zhou. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature Methods*, 17(2):193–200, 2020.

[55] V. Svensson, S. A. Teichmann, and O. Stegle. Spatialde: identification of spatially variable genes. *Nature Methods*, 15(5):343–346, 2018.

[56] H. Teng, Y. Yuan, and Z. Bar-Joseph. Clustering spatial transcriptomics data. *Bioinformatics*, 38(4):997–1004, 10 2021.

[57] F. W. Townes and B. E. Engelhardt. Nonnegative spatial factorization, 2021.

[58] F. W. Townes, S. C. Hicks, M. J. Aryee, and R. A. Irizarry. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome Biology*, 20(1):295, 2019.

[59] V. A. Traag, L. Waltman, and N. J. Van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):1–12, 2019.

[60] B. Velten, J. M. Braunger, R. Argelaguet, D. Arnol, J. Wirbel, D. Bredikhin, G. Zeller, and O. Stegle. Identifying temporal and spatial patterns of variation from multimodal data using mefisto. *Nature Methods*, 2022.

[61] W. Wang, Q. Zhong, L. Teng, N. Bhatnagar, B. Sharma, X. Zhang, W. Luther, L. P. Haynes, R. D. Burgoyne, M. Vidal, S. Volchenboum, D. E. Hill, and R. E. George. Mutations that disrupt phoxb interaction with the neuronal calcium sensor hpcal1 impede cellular differentiation in neuroblastoma. *Oncogene*, 33(25):3316–3324, 2014.

[62] F. A. Wolf, P. Angerer, and F. J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):1–5, 2018.

[63] Y. Yamamoto and P. Perron. Estimating and testing multiple structural changes in linear models using band spectral regressions. *The Econometrics Journal*, 16(3):400–429, 2013.

[64] Y. Yang, X. Shi, W. Liu, Q. Zhou, M. Chan Lau, J. Chun Tatt Lim, L. Sun, C. C. Y. Ng, J. Yeong, and J. Liu. SC-MEB: spatial clustering with hidden Markov random field using empirical Bayes. *Briefings in Bioinformatics*, 23(1), 11 2021. bbab466.

[65] Y. Yuan and Z. Bar-Joseph. GCNG: graph convolutional networks for inferring gene interaction from spatial transcriptomics data. *Genome Biology*, 21(1):1–16, 2020.

[66] L. Zappia, B. Phipson, and A. Oshlack. Splatter: simulation of single-cell RNA sequencing data. *Genome Biology*, 18(1):1–15, 2017.

[67] H. Zeng, E. H. Shen, J. G. Hohmann, S. W. Oh, A. Bernard, J. J. Royall, K. J. Glattfelder, S. M. Sunkin, J. A. Morris, A. L. Guillozet-Bongaarts, et al. Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell*, 149(2):483–496, 2012.

[68] D. Zhang, X. Liu, X. Xu, J. Xu, Z. Yi, B. Shan, and B. Liu. Hpcal1 promotes glioblastoma proliferation via activation of wnt/-catenin signalling pathway. *Journal of Cellular and Molecular Medicine*, 23(5):3108–3117, 2019.

[69] E. Zhao, M. R. Stone, X. Ren, J. Guenthoer, K. S. Smythe, T. Pulliam, S. R. Williams, C. R. Uytingco, S. E. B. Taylor, P. Nghiem, J. H. Bielas, and R. Gottardo. Spatial transcriptomics at subspot resolution with bayesspace. *Nature Biotechnology*, 2021.

[70] J. Zhu, S. Sun, and X. Zhou. Spark-x: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biology*, 22(1):184, 2021.